DOCUMENT RESUME

ED 038 697 CG 005 255

AUTHOR Jacobs, Stanley S.
TITLE Confidence-Weighting as a Scoring Technique.
INSTITUTION American Educational Research Association, Washington, D.C.; Pittsburgh Univ., Pa.
PUB DATE 2 Mar 70
NOTE 12p.; Paper presented at American Educational Research Association Convention, Minneapolis, Minnesota, March 2-6, 1970

EDRS PRICE EDRS Price MF-$0.25 HC-$0.70
DESCRIPTORS College Students, *Measurement, *Measurement Instruments, *Measurement Techniques

ABSTRACT

The study investigated the measurement technique called confidence-weighting, wherein an examinee indicates what he believes is the correct answer and also how certain he is of the correctness of his answer. It was concerned specifically with the effects of two levels of penalty on the unwarranted expression of confidence, the personality correlates of confidence-weighting and the effects on test statistics of confidence-weighting. Seventy-two subjects were administered a 130 item multiple choice test under confidence-weighting instructions. All subjects completed the California Psychological Inventory. Results included: (1) no significant effects or interaction attributable to level of penalty or sex; (2) the expression of confidence in one's objective test responses is contaminated by a general personality factor; and (3) no significant increase in test reliability as a result of confidence-weighting. (Author/TL)

# CONFIDENCE-WEIGHTING AS A SCORING TECHNIQUE

Stanley S. Jacobs

*University of Pittsburgh*

Paper presented at the annual convention of the
American Educational Research Association

Minneapolis, Minnesota

March, 1970

# CONFIDENCE-WEIGHTING AS A SCORING TECHNIQUE

Stanley S. Jacobs

*University of Pittsburgh*

The study investigated the effects of several variables on the expression of confidence in the accuracy of responses to objective test items. A final examination was administered to 72 subjects under confidence-weighting instructions (Ebel, 1965) with two levels of penalty for incorrect responses. A two-way ANOVA revealed no significant main effects or interaction attributable to level of penalty or sex. A multiple correlation of .39 was obtained between an ascendance score, based on a composite of scales from the California Psychological Inventory, and a score based on the number of incorrect responses for which maximum confidence was expressed. An ANOVA on a regression analysis resulted in a significant F ($p < .05$). Although increased penalty-level had no effect on confidence-expression, the test's reliability decreased from .85 to .39, and the correlation between conventional and weighted scores dropped from .88 to .095.

---

One of the current problems in educational and psychological measurement is the assessment of partial knowledge or degree of mastery of material tapped by objective test items.

A strategy which has been advanced as a technique for extracting additional information from objective test item responses, and as a means of increasing test reliability, is known as confidence-weighting

and is usually attributed to Ebel (1965a, 1965b). Confidence-weighting is described as

> "...a special mode of responding to objective test items, and a special mode of scoring those responses. In general terms, the examinee is asked to indicate not only what he believes to be the correct answer to a question, but also how certain he is of the correctness of his answer. When his answers are scored he receives more credit for a correct answer given confidently than for one given diffidently. But the penalty for an incorrect answer given confidently is heavy enough to discourage unwarranted pretense of confidence." (Ebel, 1965a, p. 49).

Alternative scoring procedures, such as confidence-weighting, are often regarded as relatively recent measurement innovations. This is not completely accurate. For example, the technique of confidence-weighting has a long history of psycho-physical experimentation (e.g. Henmon, 1911; Hollingworth, 1913; Trow, 1923). The technique was thoroughly investigated, both directly and indirectly, in a long line of studies in educational and psychological measurement which was interrupted in the early 1940's, probably by the diversion of research talent into activities related to World War II, (e.g. Greene, 1929; Jersild, 1929; Hevner, 1932; Melbo, 1933; Wiley and Trimble, 1936; Soderquist, 1936; Swineford, 1938; 1941; Meyer, 1939; Johnson, 1940, 1941.)

More recently, a number of studies have appeared which imply that alternative scoring procedures such as confidence-weighting may serve to make measurement more precise, (e.g. Michael, 1968). There seems to be little research aimed at identifying the relevant factors operative in situations where the respondent is given some latitude.

Several studies indicate reliable individual differences in risk-taking may be operative (Slakter, 1968a; 1968b), and that some alternative-scoring procedures result in information which differs from that obtained via conventional means (Rippey, 1968).

There appear to be at least the following assumptions made in the use of confidence-weighting:

1) Students will perform in a rational manner in a confidence-weighting situation, i.e. a high relationship exists between the possession of knowledge and a willingness to express this fact and
2) the technique is not contaminated through the introduction of extraneous variables, which may bias the situation for or against certain students, regardless of achievement.

A number of studies cited reveal these assumptions are not tenable. One may conclude that any situation in which the subject is permitted some latitude in responding, such as confidence-weighting, is moderated by factors usually extraneous to the test. This has, in fact, been amply demonstrated by Votaw (1936) and Sherriffs and Boomer (1954) in investigations concerning the effects of the correction for guessing. A number of studies (e.g., Soderquist; Wiley and Trimble; Swineford; Slakter) have posited the operation of a personality variable; several ranking indices have been suggested (Slakter, 1967), but there has been little research designed to investigate the internal validity of confidence-weighting.

The present study was undertaken to determine the effects of two levels of penalty on the unwarranted expression of confidence, the personality correlates of confidence-expression and the effects on test statistics of confidence-weighting.

## Method

### Subjects

The 72 subjects in the present study were 24 male and 48 female undergraduates, predominantly freshmen and sophmores, enrolled in an introductory course in personality and adjustment.

### Procedure

Ss were administered a 130 item multiple-choice course final examination under confidence-weighting instructions. Ss were given a general explanation of the technique; instructional paragraphs preceding the examination detailed the credit and penalty allowances, with examples.

In addition to selecting what was perceived as the most correct answer for each test item, Ss indicated their degree of confidence in their response on a three-point scale (Guess, Fairly Confident, Very Confident), with graded credit and penalty. A randomly selected 36 Ss (Group A) could earn 1 point for a correct response marked "Guess," 2 points for "Fairly Confident" and 3 points for "Very Confident." If the selected option was incorrect, Ss lost 0, 2 or 3 points, depending on the category of confidence selected. The remaining 36 Ss (Group B) were informed the penalties were 0, 4 or 6 points, again depending upon the confidence-category selected.

All Ss had completed the California Psychological Inventory (CPI) (Gough, 1957) as a part of freshman orientation. These data were obtained from University records. For each subject, standard scores on the Dominance (Do) scale, Sociability (Sy) scale, Self-acceptance (Sa) scale and Intellectual Efficiency (Ie) scale were combined as a measure of ascendance. (Crites, 1961; 1964)

The measure of confidence-expression used in the present study was defined as

$$\text{CONF} = \frac{\text{number of errors for which maximum confidence was expressed}}{\text{number of errors}} \times 100$$

## Results

Table 1 displays the CONF scores obtained from male and female Ss under the two levels of penalty. The hypothesis of no treatment, no sex and no treatment x sex effects was tested using a two-way ANOVA and retained at the .05 level. (See Table 2)

TABLE 1

CONF scores for Groups A and B by Sex

| Group (by Sex) | n | Mean | Std. Deviation |
|---|---|---|---|
| A, males | 12 | 16.15 | 7.9 |
| A, females | 24 | 10.13 | 9.5 |
| B, males | 12 | 15.43 | 10.1 |
| B, females | 24 | 17.82 | 13.8 |

TABLE 2

ANOVA Testing Effects of Magnitude of Penalty

and Sex of Subject on CONF scores

| Source | df | MS | F |
|---|---|---|---|
| Treatment (A) | 1 | 3.04 | 0.03 |
| Sex (B) | 1 | 193.67 | 1.57 |
| A x B | 1 | 282.80 | 2.29 |
| Within | 68 | 123.64 | ---- |

The relationship between the variable of ascendance, as defined in the present study, and the expression of unwarranted confidence was investigated using multiple correlation. (See Table 3) An F-test on the regression of ascendance on CONF scores resulted in an F significant at the .05 level. (See Table 4)

TABLE 3

Multiple Correlation Between "Ascendance"

Composite Scores and CONF Scores

| Variable | Multiple Correlation | F Value |
|---|---|---|
| Ie | .22 | 3.62 |
| Do | .35 | 4.69 |
| Sa | .38 | 3.71 |
| Sy | .39 | 2.93 |

TABLE 4

ANOVA for Regression of "Ascendance" Composite Scores on CONF Scores

| Source | df | MS | F |
|---|---|---|---|
| Due to regression | 4 | 331.07 | 2.93* |
| Deviation about regression | 67 | 112.87 | |

*p < .05

The split-halves reliability of unweighted and confidence-weighted achievement test scores and CONF scores of Groups A and B are summarized in Table 5.

TABLE 5

Split-half Reliability Estimates; Groups A and B

| Variable | A | B |
|---|---|---|
| Unweighted achievement scores | .89 | .79 |
| Confidence-weighted achievement scores | .87 | .39 |
| CONF scores | .86 | .68 |

## Discussion

It is apparent that, at least with naive Ss, moderate shifts in the level of penalty associated with incorrect responses do not modify the expression of unwarranted confidence. An additional analysis of the frequency of usage of the three possible confidence

categories in Groups A and B revealed no significant difference. It should be pointed out that these were naive Ss, who lacked any "baseline," based on previous experience, for their expression of confidence. Graded levels of penalty may have some effect on students with prior experience with confidence-weighting.

It appears that the expression of confidence in the accuracy of one's responses to objective test items is contaminated by a more general personality factor. The procedure seems biased against ascendant Ss, since a greater number of their incorrect responses are given with maximum confidence, which results in a greater incurred penalty.

One of the often-cited reasons for employing confidence-weighting, that of affording an increase in test reliability, finds no support in the present study.

As may be seen in Table 5, the data, with one exception, were quite reliable. Of interest is the difference in internal consistency of confidence-weighted achievement scores in Groups A and B. Apparently, the asymmetric credit-penalty relationship served to magnify inconsistencies with which confidence-weighting was employed. To clarify this point, the correlation between conventional and confidence-weighted scores was obtained and found to be .88 for Group A and .095 for Group B.

In summary, there seems to be little advantage in the use of confidence-weighting as a means of tapping partial knowledge, since it is contaminated or moderated by personality. Reasonable levels of

penalty do not effectively control the expression of unwarranted confidence, but increased penalties destroy the internal consistency of the data, and result in scores which bear no relationship to basic "number right" scores. The situation with reference to reliability gains noted with confidence-weighting is unclear; it may be hypothesized that this would occur maximally with tests of low reliability, where the introduction of an independent, reliable (but extraneous) source of variance might result in marked shifts in estimated reliability.

The time and effort required to both complete and score a confidence-weighted test are substantially greater than that required for a conventional test. The present study finds no advantage in terms of amount or precision of information.

Perhaps the best way to deal with individual differences in risk-taking propensities is to require Ss to respond to all items, even though this may increase error variance somewhat, and thereby decrease reliability. With reference to increased precision, it appears that the conventional route of item analysis and revision is less suspect.

## REFERENCES


Crites, J.O. The California Psychological Inventory: I. as a measure of the normal personality. Journal of Counseling Psychology, 1964, 11, 197-202.

_____, et. al. A factor analysis of the California Psychological Inventory. Journal of Applied Psychology, 1961, 45, 408-414.

Ebel, Robert L. Confidence-weighting and test reliability. Journal of Educational Measurement, 1965a, 2, 49-57.

_____. Measuring Educational Achievement. Englewood Cliffs: Prentice - Hall, Inc., 1965b.

Gough, H.G. California Psychological Inventory. Palo Alto: Consulting Psychologists' Press, 1957.

Greene, E.B. Achievement and confidence on true-false tests of college students. Journal of Abnormal and Social Psychology, 1929, 23, 467-78.

Henmon, V.A.C. The relation of the time of a judgment to its accuracy. Psychological Review, 1911, 18, 186-201.

Hevner, K. A method of correcting for guessing in true-false tests and empirical evidence in support of it. Journal of Social Psychology, 1932, 3, 359-362.

Hollingworth, R.L. Experimental studies in judgment. Archives of Psychology, 1913, 29, 1-119.

Jersild, A. The determinants of confidence. American Journal of Psychology, 1929, 41, 640-642.

Johnson, D.W. Confidence and achievement in eight branches of knowledge. Journal of Educational Psychology, 1941, 32, 23-26.

_____. Confidence and the expression of opinion. Journal of Social Psychology, 1940, 12, 213-220.

Melbo, I.R. How much do students guess in taking true-false examinations? Educational Method, 1933, 12, 485-87.

Meyer, G. The choice of questions on essay examinations. Journal of Educational Psychology, 1939, 30, 161-171.

Rippley, R. Probabilistic testing. Journal of Educational Measurement, 1968, 5, 211-215.

Sherriffs, A.C. and Boomer, D.S. Who is penalized by the penalty for guessing? Journal of Educational Psychology, 1954, 45, 91-90.

Slakter, M.J. Risk-taking on objective examinations. American Educational Research Journal, 1967, 4, 31-43.

Slakter, M.J. The effect of guessing strategy on objective test scores. Journal of Educational Measurement, 1968a, 5, 217-221.

Slakter, M.J. The penalty for not guessing. Journal of Educational Measurement, 1968b, 5, 141-144.

Soderquist, H.O. A new method of weighting scores in a true-false test. Journal of Educational Research, 1936, 30, 290-292.

Swineford, F. Analysis of a personality trait. Journal of Educational Psychology, 1941, 32, 438-444.

_____. The measurement of a personality trait. Journal of Educational Psychology, 1938, 29, 289-292.

Trow, W.C. The psychology of confidence - an experimental inquiry. Archives of Psychology, 1923, 67, 1-47.

Wiley, L.N., and Trimble, O.C. The ordinary objective test as a possible criterion of certain personality traits. School and Society, 1936, 43, 446-448.

Votaw, D.F. The effect of do-not-guess directions upon the validity of true-false or multiple-choice tests. Journal of Educational Psychology, 1936, 28, 698-703.